信息检索研究现状简述

王斌 李鹏

摘要：近年来，一方面，信息检索在自身发展的同时不断和其他学科领域交叉融合，另一方面，新资源、新平台的出现也促进了信息检索的迅猛发展。信息检索研究呈现出个性化、协同化、社会化的趋势。本文总结了近年来信息检索研究的一些新动向，并分析了未来发展的若干趋势。

关键词：信息检索；个性化；协同搜索；社会化搜索

1　引言

信息检索(Information Retrieval, IR)是研究如何从大规模原始信息中快速准确全面地获取用户所需信息的一门学科。它最初起源于图书馆的文献查找需求，后来扩展到各种信息处理领域。互联网的出现、全球数字化进程的加快，使得信息过载(Information Overload)问题日益严重。从大量信息中找到符合用户要求的信息已经成为非常迫切的需求，同时也是一个挑战。这促进了信息检索技术的发展。特别是最近几年来，随着通用搜索引擎的日益流行、各行业搜索引擎的涌现、各商家对搜索技术的日益重视和大量投入，作为搜索引擎核心的信息检索技术的研究也出现了一个前所未有的高潮。原来研究自然语言处理、人工智能、机器学习、统计、认知科学、数据库、分布式并行处理的不少学者都将目光投向了信息检索这个历久弥新的应用。信息检索已经成为一门跨学科跨领域的交叉学科。以顶级学术国际会议为例，原来只有 SIGIR 等为数不多的会议收录信息检索相关的最新成果，现在包括 SIGKDD、ICDM、WWW、SIGMOD、NIPS、VLDB、ACL、IJCAI、AAAI、EMNLP、CIKM 等等在内的各领域的顶级会议都收入了不少有关信息检索研究的论文，有些会议信息检索相关的论文甚至占到绝大部分。可以说，信息检索的研究进入到一个前所未有的各种技术交叉融合的时代。本文中，我们对近年来信息检索相关的研究动向进行了梳理和总结，以期能够为相关研究人员提供参考。

任何一个信息检索系统都不外乎如下的结构：

用户将自己的需求(Information Need)表达成查询(Query)提交给检索系统，检索系统从文档集合(Collection)中对每篇文档(Document)和查询进行某种相似度计算(Similarity Computation)，从中输出部分可能满足用户需求的结果(Result Set)。不同的信息检索应用可能在查询、文档集、结果集合以及相似度计算的要求上不尽相同，从而出现了各种不同的应用。最常见的信息检索系统包括以万维网（Web）搜索引擎为代表的信息搜索系统、以信息订阅系统为代表的信息推荐系统以及以回答结果为目标的问答系统等等。

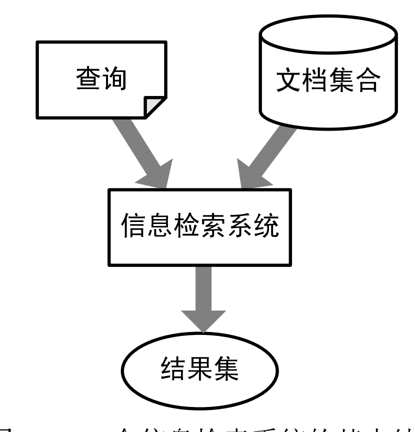


图 1. 一个信息检索系统的基本结构

为介绍方便，我们把整个信息检索相关的研究归结成 3 个层次(如图 2 表示)，从底往上

分别是资源层、技术层和应用层。资源层主要包括信息检索研究利用和使用的资源，除了人们常用到的 TREC[1]数据、WordNet[2]、HowNet[3]、网上其他公开语料等数据外，我们在图中主要列举了近年来在研究中广泛采用的资源或工具。像社交网络(如 Facebook)数据、微博(如 Twitter、新浪微博)数据、百科(如维基百科、互动百科)数据等等。第二层主要是技术层，包括自然语言处理(NLP)、机器学习(ML)、数据挖掘(DM)、并行处理(PP)在内的多项来自不同领域的技术。它们和传统的信息检索技术(建模、索引、查询扩展、相关反馈等)相融合。第三层是应用层，包含了大量基于信息检索技术的应用，如网页搜索引擎、微博（Microblog）搜索、博客（Blog）搜索、论坛搜索等等。另外，需要特别一提的是信息检索中的评价技术。众多技术、众多应用的出现使得评价愈发重要，评价是信息检索领域的一个永恒话题。

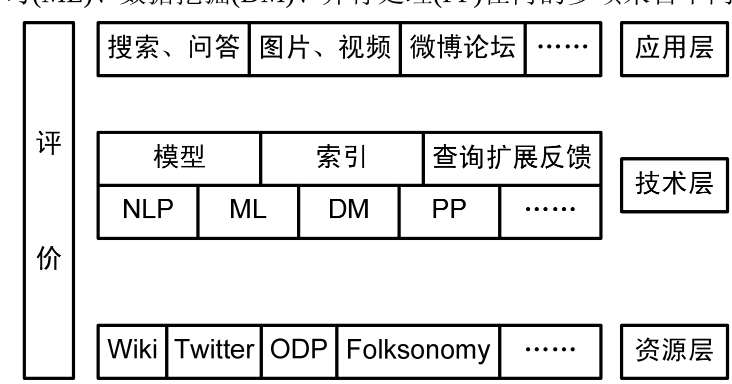


图 2. 信息检索研究层次图

2 信息检索新应用的研究

在信息检索的相关研究中，针对新应用的研究近年来占据主要地位。一方面，网上不同格式、不同形式、不同领域信息的日益增多使得针对这些信息进行检索的需求日益强烈；另一方面，新信息类型的出现使得针对这些信息的检索呈现出新的各自不同的要求，从而产生特定的研究问题。以下列举一些有代表性的应用研究。

（1）. 多媒体检索研究

传统的信息检索研究主要针对文本对象。随着多媒体文档的日益增多，对多媒体检索的需求也越来越强烈。根据媒体对象的不同，多媒体检索又可以分成图像检索、视频检索、语音检索、音乐检索等等不同类型。这些不同的研究也形成了自己的研究社区。如：国际上著名的图像视频检索会议 CIVR(ACM Conference on Image and Video Retrieval)自 2002 年开始已经召开了 10 届。针对视频检索的国际著名评测会议 TRECVID(http://www-nlpir.nist.gov /projects/trecvid/)也召开了多届。研究音乐检索的研究人员组成了国际音乐检索协会(The International Society for Music Information Retrieval, http://www.ismir.net/)，并从 2000 开始组织了 12 届年会。

多媒体检索可以看成是媒体处理和传统信息检索技术的综合应用。通过媒体的分析和处理，可以得到媒体的低级特征(如：颜色、形状、纹理、音量等)。然而，用户的查询往往是高级的语义概念(如：日出、五星红旗等等)，媒体的低级特征到高级的语义概念之间存在着语义鸿沟(Semantic Gap)。可以说，多媒体检索的最终目标就是要解决如何跨越语义鸿沟这

[1] TREC(Text REtrieval Conference)是信息检索领域的一个著名评测会议，它由美国标准技术研究所组织举办，以期达到在大规模共同数据平台上对信息检索技术进行评价的目的。TREC 提供数据也成了研究人员的常用标准实验数据之一。详细信息参见：http://trec.nist.gov。

[2] 由普林斯顿大学心理学家、语言学家和计算机工程师联合设计并开发的一部英语词典，它包含英语单词之间的语义关联，目前在学术界使用广泛。详细信息参见：http://wordnet/princeton.edu。

[3]由董振东、董强设计并实现的一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，目前在学术界使用广泛。详细信息参见：http://www.keenage.com。

个核心问题。近些年的研究表明，综合利用各种媒体或模态信息(如同时在视频中的音频流、视频流和文本流)来提高语义映射的精度看上去是一条可行之路。本质上说，这些研究实际上在建立不同媒体表达形式之间的关联。在此基础上，有人提出了跨媒体检索(Cross Media IR)研究。可以从两方面去理解这类研究：一方面，输入某种媒体形式的查询，可以在不同媒体形式的文档集合中进行检索；另一方面，可以综合多种媒体或模态信息来提高检索结果的质量。

目前，商用的多媒体搜索系统主要还是基于文字标注的系统，要实现将媒体处理技术融入检索技术的成熟应用系统还需要较长的一段路要走。

（2）． 针对不同领域或不同场景的信息检索研究

人们习惯于将不同于通用搜索的系统称为“垂直搜索”系统。所谓“垂直”可以理解为针对不同领域不同场景进行信息检索。区别于传统通用搜索的研究，“垂直化”搜索往往要求更高的数据质量。因此，这些研究中信息抽取(Information Extraction)，即从文档中分离出所需要信息的技术，往往是其中的关键技术。比如，针对人物的搜索要求能够从文档中分析出人物的各种属性信息(如：性别、年龄、工作等等)；针对产品的搜索要求得到产品的各种属性(如：型号、价格、产地等等)。如果抽取的目标文档格式化程度较高，可以通过规则模板或机器学习的方法进行分类。如果抽取的目标文档格式化程度不高，往往还要用到文本理解技术。如果信息分散在不同目标中，还需要利用信息集成(Information Integration)技术。信息抽取的质量往往决定了最后“垂直搜索”的效果。在实际中，通常是通过人工+计算机处理的方法来进行处理。针对不同的领域，这些年还出现了针对医学文献、生物文献、专利文献、法律文献等的检索研究。比如：TREC 会议从 2003 增加基因相关文献的检索(有人认为这种研究也是“生物信息学”的一种) ；从 2006 年开始增加了法律文献的检索。另一个重要评测会议 NTCIR(http://research.nii.ac.jp/ntcir/)从 2002 年开始就增加了专利文献的检索。这些研究集中关注该领域的应用需求，比如医学文献检索中可能需要融入药物和病症之间的关联分析。

另外，新的网络事物的出现也会促发信息检索研究。最典型的例子是前些年出现的博客（Blog）检索和近两年出现的微博(Microblog)搜索。作为 WEB2.0 时代的产物，博客受到人们的广泛关注，分析博客从而进行检索的需求非常强烈。不同于以往的检索对象，博客有它自己的一些特性，如：博客包含博主、博文、评论、链接、引用通告(Traceback)等信息。总的来说，它是各种信息的一个聚合体。针对博客这个新鲜事物的研究，比如对垃圾博客的识别和过滤、博客社区的发现和分析、博客内容挖掘及趋势预测，这几年比较活跃。更值得一提的是，由于博客上存在大量的主观评论，因此，博客评论的倾向性分析(Sentiment Analysis)日益受到关注。很显然，这种技术也能广泛应用到商品的评论中，从而具有很要的商业价值。简单地说，倾向性分析就是指对文本是否存在主观评论、评论的对象是什么、是褒还是贬、程度如何等进行分析。有些人也称之为情感分析或者观点分析。TREC 还从 2006 年开始增加了博客检索的子任务，要求返回的文档不仅要考虑相关性还要考虑是否存在倾向性。由于倾向性分析涉及到文本处理技术，也引起了自然语言处理领域研究人员的广泛兴趣。国内外出现了大量针对该问题进行的研究。国内中文信息学会还组织了中文的倾向性分析评测。从目前的研究结果来看，大部分工作仍然基于词汇层面(情感词)，这项技术要真正走向实用还要付出艰苦的努力。

微博由于其和移动互联网的紧密结合而在最近几年蓬勃发展起来。微博信息短、传递快、实时性强，包含用户信息、不同用户之间的关注信息、信息的转发路径等信息。有研究指出，微博搜索在查询意图、数据分布等诸多方面不同于传统的搜索，是一项有前途的研究方向。

（3）． 移动搜索(Mobile Search)研究

移动设备(以手机为代表)的高普及率、手机网络的高覆盖率和发展前景、手机和用户的紧密绑定关系、手机的庞大用户群等等因素，无疑使基于手机的信息搜索具有重要的商业价值。WWW、SIGIR 等一系列顶级检索会议都多次举行了移动搜索的研讨会。和普通搜索一样，手机搜索同样要对信息进行获取、组织和提供访问。不同的是，目前手机搜索研究的基本出发点是突破手机输出(主要指屏幕显示)和输入的限制。由于手机屏幕尺寸的限制，一方面要求返回的搜索结果更精确，尽量杜绝垃圾信息；另一方面，也要求在有限的屏幕空间下结果的布局更合理，显示更简洁，显示重点更突出，便于用户进行进一步浏览操作。这需要综合排序算法、信息过滤、文本分析、摘要、人机交互等各种技术。而由于在手机上用户输入的限制，在检索交互上，往往要通过拼音文字转换、查询推荐、查询补全等技术来尽量减少用户的输入负担。当然，手机搜索中除了文本搜索，多媒体搜索也是一个重要组成部分。媒体的标注、显示、传输都是手机多媒体搜索中主要的研究问题。另外，由于手机本身的特点，可以考虑把用户因素、地理位置等下上文环境(Context)因素考虑在内进行搜索结果的优化研究(如进行个性化搜索—Personalized Search 或者本地搜索—Local Search)。

目前移动搜索的研究仍然刚刚起步。由于各方面(网速、上网费、手机功能等)的限制，现在移动搜索的普及率远不能和互联网上搜索引擎相比。但是可以预见，在不久的将来，移动搜索必将有更广阔的前景。

（4）． 基于检索的广告技术

现有的大部分商业搜索引擎都有在线广告(Online advertisement) 。用户输入点击，在搜索结果中或侧部会出现可能与用户查询相关联的广告。在线广告是搜索引擎公司巨大收益的主要来源，也能给广告源商家带来重要利益，因此，受到商业界和研究界的高度关注，甚至衍生了一个叫计算广告学(Computational Advertising)的新名词。SIGIR、WWW 等会议最近几年也把信息检索中的广告作为一个重要的议题进行研讨。输入的可以是一个搜索查询，往往称为付费搜索（paid search），匹配得到的广告称为赞助广告（Sponsored ads）或关键词驱动广告（keyword-driven ads)；也可以是用户正在浏览的网页、图像或视频，得到的广告称为 context-driven ads(语境驱动广告)或 contextual ads（语境广告）。然后，系统根据从用户的输入中计算出来的意图将合适的广告推送给用户。从刚才的过程可以看出，计算广告非常像信息检索，可以看成是根据用户的输入查询在广告库这个集合中进行匹配，将最可能的广告推送给用户。信息检索技术很显然能在计算广告中发挥巨大作用，这也是为什么在计算广告在信息检索被广泛关注的主要原因。与一般信息检索不同的是，计算广告中的匹配并不简单地基于传统的相关度概念，而是要挖掘用户的商业意图。举例来说，一个用户输入“北京到上海 车票”，系统可以猜测用户很可能要到上海，于是可以把上海的一些酒店的广告信息推送给用户。这里就不是简单的相关度匹配的概念。如果没有从用户输入中分析到商业意图，系统也可以不推送广告。计算广告要同时考虑匹配结果的准确性和用户的良好体验。在排序时，计算广告中要综合各种利益，要同时考虑用户的体验、商家的利益以及商业上的限制等等。因此，这里的排序是一个非常复杂的问题。除了原有的排序算法，往往还同时使用经济收益模型。另外，广告文本一般都不会很长，缺乏像传统文本中可以利用的上下文语言信息。因此，传统的信息检索模型也需要做相应修改。总之，计算广告学是一门新兴的具有挑战性的研究领域。当然，目前进入这个领域还有相当的难度，而缺乏数据是一个瓶颈。广告数据涉及商家利益，不便公开。虽然，微软曾经宣称要提供数据来支持该领域的研究，但是目前仍未实现。对结果的评估是另外一个主要问题。因为，计算广告最终是要让各利益攸关方的利益最大化：用户体验最好、搜索引擎商家和广告商家利益最大化，这种目标难以在研究中

进行模拟。因此，对研究结果的有效性评估是值得研究的一个重要问题。

（5）． 个人信息管理(Personal Information Management, PIM)及桌面检索技术研究

在很多研究者面向巨大的互联网资源进行研究的同时，人们发现，由于硬盘的价格越来越便宜，用户自身拥有的数据越来越多，每个人的机器已经不堪数据重负。个人信息管理已经成为一个非常重要而迫切的需求，逐渐开始受到研究界的广泛重视。2005 年开始，国际上就召开了一年一度的个人信息管理国际研讨会，SIGIR、SIGCHI(人机交互)等会议也在这几年把个人信息管理列为主要讨论议题。国际顶级期刊 ACM Transaction on Information System (TOIS)也在 2008 年年底组织了一期个人信息管理的专刊。个人信息管理主要研究个人信息的获取、组织、管理、维护和检索，也是兵家必争的“未来桌面”的核心技术。个人信息管理中我们比较熟悉的是桌面搜索。近年来，桌面检索(对用户硬盘上的数据进行搜索)技术已成了搜索引擎公司关注的焦点，很多公司都开发了自己的桌面搜索引擎。但是，商用的桌面搜索工具在用户满意度方面仍然有待提高。一方面，同传统的网页搜索相比，桌面搜索可以利用的信息很少。网页搜索对网页的检索除了利用关键词外，还可以利用网页之间的链接关系、用户日志等信息。而桌面搜索最初只有文本信息和访问时间等信息。另一方面，用户对桌面搜索的准确度要求却比网页搜索高，通常用户希望能通过桌面搜索引擎直接定位到想要的文件。所以桌面搜索中如何获得更多的信息，并利用这些信息对检索结果进行排序是一个很困难的问题。当然，桌面搜索也有一定的优势，比如和用户结合得很紧，可以利用用户的信息。我们相信，结合用户行为信息来做桌面搜索是一个有前途和可行的研究方向。目前，已经有文献利用用户访问文件和查询的日志，使用机器学习的方法，学习适合于个人的排序算法。另外，也有人利用访问模式来建立文件之间联系方法，然后利用类似 PageRank 的算法进行排序。

（6）． 社会化挖掘及搜索(Social Mining and Search)研究

近年来，互联网越来越呈现出明显的社会化趋势。以 Delicious 等为代表的社会化标签网站积累了大量数据，以 Facebook 为代表的社会化网络得到网民的广泛参与，以 Twitter、新浪微博为代表的新的信息共享和传递机制表现出勃勃生机。这些新生的事物一方面促进了一些新的搜索应用的出现，如 Twitter 搜索、微博搜索等，另一方面由于其蕴含了大量社会化信息(用户信息、用户关系信息、用户行为信息、信息关联信息等)而为其他应用提供了十分宝贵的数据。研究人员正在挖掘这些数据背后隐藏的规律和深刻内涵，来进一步提高信息检索的效果。

3 信息检索技术的研究

这一部分我们介绍信息检索相关技术的研究。传统的信息检索技术主要包括信息检索模型、相关反馈和查询扩展、索引技术等等。最近一些年来，机器学习、自然语言处理、统计等其他领域的技术被更广泛地用于信息检索中，进展情况小结如下：

（1）． 信息检索模型的研究

检索模型的本质是对用户需求和文档的相关性建模，主要包括查询和文档的表示技术及相关性排序技术。早期的布尔模型、向量空间模型、概率模型及 1998 年出现的统计语言建模检索模型仍然在被人们不断改进和应用。此外，也陆续出现了几种新的模型。主要的改进思路在两个方面，一个是如何将特征项之间的关系考虑在内来突破传统模型中的特征项独立(Term Independence)假设；另一个是如何突破传统的词项频率 TF(Term Frequencey)、逆文档

频率 IDF(Inverse Document Frequency)及文档长度等三个因素来改进检索模型。马尔科夫随机场(Markov Random Field，MRF)模型可以看成是前者的一个结果。它综合考虑了特征项之间的各种组合关系，并通过不同权重融合到一个检索模型中。实验结果表明，马尔科夫随机场模型能够取得超过传统检索模型的结果，在有噪音的数据(如网页)上效果更加明显。传统的信息检索模型，不论是具有 30、40 年历史的向量空间模型、概率检索模型还是近 10 年出现的语言建模模型，都只包含了三个因素：词项频率、逆文档频率及文档长度，更多因素的融入一直是人们的研究目标。有人提出将特征项之间的邻近关系(Proximity)也引入到信息检索模型中，并进行了初步尝试。关于信息检索模型的一个更有意思的研究来自伊利诺伊大学厄本那-香槟分校（UIUC）。他们提出：一个好的信息检索模型必须满足一些基本约束条件，并证明现有的信息检索模型条件满足(和参数有关)或不完全满足上述基本条件，在传统模型上和参数变化相关的实验结果印证了其理论分析。在此基础上，他们提出了构造新的检索模型的框架和方法，新提出的模型具有一定的优势。

（2）． 基于机器学习的信息检索模型研究

近年来，基于机器学习的信息检索模型研究掀起一股热潮。排序学习（Learning to Rank[4]）这个议题在 SIGIR 论文中占据了不小的篇幅；一些机器学习的会议(如 NIPS)也纳入了这个议题；微软亚洲研究院还专门建立了相应网站(http://research.microsoft.com/en-us/um/beijing/projects/letor/index.html)，提供相关论文、数据、评测标准和工具，供研究者使用。与现有的启发式排序函数不同，这些研究假定排序函数满足某种形式，然后通过在标注集合上进行训练的方法求出模型参数，从而得到最后的排序函数。排序学习巧妙地将排序问题转化成机器学习问题，因而受到研究人员特别是从事机器学习的研究人员的特别关注。各种机器学习的方法被引入到检索当中，包括有监督的学习（Supervised Learning）和半监督的学习(Semi-supervised Learning)、生成式(Generative)机器学习方法和判别式(Discriminative)机器学习方法。各种传统的机器学习方法也被“改装”成适合于排序学习的方法，如 Ranking SVM、RankBoost、RankNet 等等就是这些年提出的方法。尽管还存在各种争议，理论上也有待完善，但是排序学习正日益受到广泛的关注是不争的事实。更重要的是，它使得机器学习的研究人员能够很快地参与到信息检索的研究当中，从而为信息检索的研究增加了生力军。

（3）． 查询分析技术的研究

在一个典型的信息检索系统当中，用户将自己的信息需求表示成查询输给检索系统。检索系统根据查询将结果按照匹配的相关程度高低返回给用户。这其中，查询是用户和计算机之间的交互“语言”，起着承上启下的枢纽作用：一方面，它要尽可能贴切地反映出用户的信息需求；另一方面，输入的查询要能被检索系统所理解和处理。然而，由于多种原因(如用户背景、经验的差别)，使得用户输入的查询不能贴切地反映用户的信息需求，初次查询不一定能返回满足其需求的结果。而传统的信息检索往往基于关键词匹配，不能理解查询背后所隐藏的深刻意图(比如：输入“字处理 共享 软件 下载”的用户希望能够得到一个能够快速下载相关软件的网站，而很多搜索引擎只提供关键词匹配的无关结果)。因此，需要对用户的查询进行深刻的分析和理解，以便能够对原始查询进行重构或者针对用户的意图来有针对性地检索，以提高信息检索的精度。其最终目的是减少用户到达目标文档的时间。近些年，针对用户查询分析的研究如雨后春笋般涌现出来，在一系列重要会议上都占据了较大的篇幅。

在查询分析中意图分类有重要地位。意图的分类体系很多，比如有人将查询分成信息类

[4] 亦有译作“学习排序”

(Informational)、导航类(Navigational)和事务类(Transactional) 。简单地说，信息类就是查找与查询相关的各方面的信息，比如“中国 历史”，希望返回与之相关的各方面信息；导航类主要针对某个网站入口或者个人主页进行查询，比如“新浪 首页” ；事务类的查询得到结果以后通常还要进行后续的交互操作，比如输入“字处理 共享 软件 下载”得到检索结果以后还要进行下载等操作。分成这几类查询以后就可以在检索中采用有针对性的检索方法。比如有人经研究发现，导航类查询检索中锚文本的作用会很大。因此，一旦判定是此类查询便可以加大锚文本的权重，从而返回更好的结果。除了上述分类体系外，查询也可以按照领域进行语义分类，如分成“计算机”、“物理”、“化学”等领域类别。有人根据查询的意图中是否包含商业意图进行分类；也有人根据输入的查询意图是否具有歧义进行分类；还有人根据查询意图是否包含多媒体需求进行分类。由于查询通常都较短，本身提供的信息量不足，所以对查询进行意图分类，往往都需要借用外部资源，比如通过查询日志进行训练，归结出分类规则。

对查询进行分析的另一种技术是查询难度预测。这通常是通过查询和返回的检索结果来判断结果的优劣。如果结果较好则对应“易”查询，结果较差则对应“难”查询。将查询区分成“难”和“易”有很多应用前景，比如：一旦用户输入“难”查询时，我们可以将与之相关的“易”查询推荐给用户，以便获得较好的结果。近几年，查询难度预测的研究非常热。很多人在 SIGIR、CIKM 上发表了大量与之相关的文章。这些研究主要是从结果分布的特点来判断结果的好坏。

另外，还有一些研究通过分析查询的其他特性，进行有针对性的检索研究。比如，查询是否需要个性化、查询是否可本地化等等。

4 新资源在信息检索中的利用

这一部分我们将介绍新的资源在信息检索中的利用。新的资源的兴起不仅对传统的资源提供了补充，还由于其特点而被信息检索研究所利用。

（1）. 维基百科（Wikipedia）:

维基百科，是一个开放式的网络百科全书。其自由、免费、内容开放的百科全书协作计划吸引了来自世界各地的参与者，目前已经成长为全球最大的网络百科全书。截至 2008 年 1 月，英文版维基百科已有 6,000,000 多个条目，并且还将不断增加。近年来，维基百科以其数据量大、质量高、协同编辑等特征脱颖而出，成为信息检索和自然语言处理领域的研究热点之一。在维基百科中，每一个条目都对应一篇文章，并且这个条目是对这篇文章所描述事物的概括，或者说是主题。 这些主题大部分由词组构成。由于维基百科允许网络用户在一定规范内自由编撰，所以这些主题更能反映现实生活中人们常用的语义概念。随着维基百科不断地被编辑和增加条目，我们可以期待其包含的人们常用的词组将更为完善，同时能及时反映人们用语的变化。维基百科为解决自然语言理解和信息检索中的问题提供了新的资源和方法。在维基百科中，概念之间存在包括上下位在内的多种关系。因此，一种直接的应用是将维基百科作为词典，通过建立概念之间的关系图，来计算概念之间的语义相似度。有人采用维基百科来解决命名实体消歧的问题，其作用在于将原有的文字片断拓展，根据维基百科的内容提供消歧所必需的上下文信息。很多研究利用维基百科改善检索的结果

（2）. 开放式目录管理（Open Directory Project，ODP）：

1998 年 6 月，当时一位程序员里奇.斯克伦塔（Rich Skrenta）对雅虎的搜索结果中经常

出现老的和死的链接感到非常厌烦，于是他在互联网上发出了倡议，请求位于全球各地的互联网用户都志愿来帮助编辑这个目录。倡议很快得到了很多热心志愿者的支持，于是划时代的管理方式开放式目录管理 ODP 就此诞生。国外最著名的开放式目录管理网站当属 Dmoz（http://www.dmoz.org），一般大家所说的开放式目录管理就是指的该网站。Dmoz 由超过 8 万名编辑志愿义务工作，将 4 百多万个网站分类到 59 万多个详细类别中，所有信息提供给任何个人和组织免费使用。开放式目录管理中包含了大量分类目录信息，而且都是人工编辑添加并经过相应专业管理员处理的数据。因此很多公司和研究人员利用开放式目录管理提高检索、分类和聚类效果。例如，谷歌（Google）在搜索结果排名中就考虑了网站在开放式目录管理中的信息；很多人利用开放式目录管理中的分类目录信息取得了不错的效果，如：帕罗.费拉吉那（Paolo Ferragina）等利用开放式目录管理对搜索结果进行聚类并提高了用户搜索体验[5]。因此，这引起了研究人员的关注，KDD-CUP 2005[6]就有一个将 80 万个查询分到 67 个类别的任务。

（3）． Folksonomy（分众分类法）：

Folksonomy 这个词由 folks（人众）与 taxonomy（分类学）组合而来，也有人译作社会分类法，是指一种由用户对 web 资源（网页、图片等）标注，进而集合大众对某个资源的标注来对该资源分类的协同工作方式。在这种模式下，用户既是标签(tag)的使用者，同时也是创造者。用户标注的标签反映了用户对于资源的认知，而不同的用户对同一资源标注不同的标签，则从不同方面放映了该资源的属性。在 web2.0 时代，web 提供了对网页、图片等进行标注的机制，互联网用户可以方便地对浏览的信息进行标注。Folksonomy 的典型系统有：http://delicious.com/（分享书签的网站），http://flickr.com(分享照片的网站)等。通过 delicious.com，用户可以保存自己喜欢的网页，同时根据自己的相关性判断对网页标注标签，有同样兴趣的用户即可通过标签查找到网页并进行浏览。Folksonomy 有很多好处，如：允许本体（ontology）、辞典及分类系统的发展，可以搜索及方便地浏览，可以发现新事物，发现社区，进行协作推荐等。同时，Folksonomy 也有难点，如各种语言标签的混合，标签的单复数、歧义、同义及抽象具体的程度难以控制，垃圾标签的干扰等。一些学者也在利用自然语言处理等技术开展研究，力图解决这方面的问题。

标签对信息的组织及检索有重要的作用。在文本检索中，标签可以加入到向量空间模型中，参与文档的表示；也有做法使用标签信息表示文档之间的关系，从而协助文档的排序。标签对于非文本信息的组织及检索也有重要的意义。通过对图片、视频等多媒体资源的标注，可以对这些非文本信息进行组织、归类及检索。同时，标签信息还可以帮助用户进行知识管理，跟踪事物发展，发现新资源，找到志同道合者等。

（4）． 搜索日志（Search Log）：

搜索引擎会对用户的搜索行为进行记录，形成大量的日志。一般情况下，搜索日志中会包含用户 ID、查询词、结果列表以及点击情况等信息。在搜索日志的研究中，有很大一部分是做一些统计分析工作。比如为了了解用户的搜索习惯，可以统计查询词的长短、一次查询之后的点击次数、查询在各个领域里的分布、用户为了得到一个想要的信息平均进行的查询次数、查询词出现的频率（热门程度）、新出现的查询词，等等。这些统计分析工作对了解用户的搜索习惯、掌握舆情、了解社会的热点等等有很大的帮助。随着研究的深入，有很多研究者开始利用搜索日志来帮助改进搜索引擎本身，还有的研究者把它作为一种新的数据

[5] Paolo Ferragina and Antonio Gulli, *A personalized search engine based on web-snippet hierarchical clustering*, WWW 2005

[6] 一个定期举行的知识发现和数据挖掘竞赛

资源从中发现有价值的数据。比如搜狗输入法的词库，就使用了从用户的搜索日志中发现的新词；而在搜索日志上进行命名实体识别、查询的理解、挖掘查询词之间的相似度来实现查询推荐等等的研究也越来越多。

5 信息检索的评价研究

评价问题一直是检索领域的基础性问题，涉及到检索的各个方面，最主要的研究工作集中在检索有效性上面。检索有效性是根据返回结果中相关文档所排的序，对检索系统的性能给出评价。评价涉及到评价过程和评价指标。

（1）． 评价过程研究：

评价过程的目标是获得相关性判断。传统的相关性判断是通过人工标注来完成。由于进行检索的语料集规模比较大，所以对于某个特定的查询，人工来完全标注相关文档是无法接受的。汇聚（pooling）方法提供了一种解决途径。汇聚的基本思路是：通过将每个系统提交的前面的若干条检索记录进行求并运算，去掉重复的文档，构成要判断的文档集合。该集合中的文档被认为是所有的相关文档，未进入该集合的文档默认为是不相关的。可以看出，这种方法并不能标注出所有的相关文档。但是通过对评价结果的分析，这样的评价过程对结果的影响不大，所以在 TREC 中得到了广泛的应用。

近些年来，TREC 中有关检索有效性评价的语料库变得越来越大，terabye track 和 million query track 使用的 GOV2 语料库大小为 426G，包含文档 25205197 篇，对于每个话题（topic），一般选前 10000(terabyte track)或前 1000(million-query track)篇文档来评价检索性能，使用的话题个数为 50 个(对 terabyte track)或 1700 个(对 million-query track)。通过简单的计算可以看出，即使是使用汇聚方法，要进行判断的文档数目也很庞大，因为汇聚方法中进行判断的文档之间没有区别，需要完全判断。能不能尽可能地减少人工标注量呢？答案是肯定的。这方面的进展要归功于本（Ben）和贾维德（Javed）的工作。评价的目标本质上是获得系统检索性能的相关性排序，如果可以得到与原始评价一致的系统排序结果，那么评价就是有效的。由于平均准确率（Average Precision， AP）的计算与相关文档出现的位置有关系，排在前面的相关文档对最终指标的计算贡献最大，对于系统间比较获得差异也最大，排在后面的文档对系统比较产生的作用相对较小。本提出按照文档对系统比较的贡献来对文档进行排序，按照顺序对文档相关性进行标注。标注到一定程度可以停止。本通过对平均准确率公式的改写，将文档的相关性看作一个随机变量，那么平均准确率也是一个随机变量。对于已经判断的文档，相关性的值是确定的。对于未判断的文档，本设计了一个基于序回归的模型。文档相关性概率可以通过在已判断的文档上建立的模型预测得到。这样可以算出每个系统在每个话题上的平均准确率期望值，进而可以得到系统的平均准确率（Mean Average Precision，MAP）的期望值。系统按照 MAP 期望值进行排序。由于只进行部分判断，该方法可以明显减少标注量；另外设计合理的预测模型，使预测得到的文档相关性概率值对最后的评价排序可以非常鲁棒。贾维德的方法则是使用抽样的思路，获得样本来估计平均准确率值。该方法也可以大幅度地减少标注量。他们的方法已经被用在了 million query 任务中。

（2）． 评价指标的研究：、

在评价中，一般是对于每个话题计算一个评价指标。系统检索性能的评价通过在多个话题上取平均来获得。常用的评价指标有 AP、R-precision、b-pref、NDCG、inferred AP 等。这些指标分别用来度量检索效果的不同方面，其中 b-pref 和 inferred AP 从 terabyte track 中引入，目的是在相关性判断减少的情况下保证评价结果尽可能与完全判断一致。评价指标一

般希望有好的数学含义，同时可以使比较结果稳定、鲁棒，即不易受到相关性判度改变的影响。

由于篇幅和能力所限，上面只简单列举了信息检索研究的部分现状。需要指出的是，在大量技术被引入信息检索领域的同时，信息检索技术也已经逐渐成为一项基础技术，被研究者引入其他领域。如有人把信息检索中的 PageRank 技术引入到软件工程领域，取得了令人瞩目的成果。

6 小结

最后，我们简单地对当前的信息检索研究进行总结，当前信息检索研究存在着几个基本特点：

（1）. 以用户为中心、以提高用户交互体验为目标。信息检索应用的最终目标是满足用户的需求，所以必须以此作为研究的最终目标。现代信息检索研究中更强调用户的中心地位，并以此驱动技术研究，可以说现代信息检索呈现出个性化的趋势。

（2）. 集中众人智慧(Crowd of Wisdom)。从资源来看，不论是微博、维基百科、开放式目录管理还是搜索日志，都集中了大量用户的智慧。从这些数据中，可以提取用户的共性，来提高信息检索的结果精度。从研究方法上看，基于用户协同(Collaboration)也就是利用用户相似性的方法不论是在过滤还是在检索上都被视为最重要的手段之一。

（3）. 新的资源、新的平台催生了一系列新的研究点。比如，微博挖掘和搜索、评论倾向性分析、移动搜索、广告推荐等等。这些新的研究点不仅具有极大的实际应用价值，也有很大的挑战性。它们大大丰富了信息检索的研究。依托社会化资源的社会化搜索正广泛受人瞩目。

（4）. 以大规模数据的分析和学习作为主要手段。现代信息检索研究中，更强调用户的中心地位，大规模的数据分析和学习是必不可少的手段。目前的信息检索研究融入了来自各领域的技术，也吸引了来自各领域的研究人员。可以预见，下一步信息检索的研究会更加丰富多彩。

作者简介：
王 斌： 中国科学院计算技术研究所、副研究员、博士生导师 wangbin@ict.ac.cn
李 鹏： 中国科学院计算技术研究所、博士生